Catch Me If You Can: How Geo-indistinguishability Affects Utility in Mobility-based Geographic Datasets

Adriano Di Luzio
Sapienza University of Rome
diluzio@di.uniroma1.it

Aline Carneiro Viana
INRIA Saclay
aline.viana@inria.fr

Konstantinos Chatzikokolakis
University of Athens
kostas@di.uoa.gr

Georgi Dikov*
TomTom, Amsterdam
gvdikov@gmail.com

Catuscia Palamidessi
LIX, École Polytechnique
catuscia@lix.polytechnique.fr

Julinda Stefa
Sapienza University of Rome
stefa@di.uniroma1.it

ABSTRACT

This paper sheds light on the trade-offs between privacy and utility in mobility-based geographic datasets. We aim at finding out whether it is possible to protect the privacy of the users in a dataset while, at the same time, maintaining intact the utility of the information that it contains. In particular, we focus on geo-indistinguishability as a privacy-preserving sanitization methodology, and we evaluate its effects on the utility of the Geolife dataset. We test the sanitized dataset in two real world scenarios: 1. Deploying an infrastructure of WiFi hotspots to offload the mobile traffic of users living, working, or commuting in a wide geographic area; 2. Simulating the spreading of a gossip-based epidemic as the outcome of a device-to-device communication protocol. We show the extent to which the current geo-indistinguishability techniques trade privacy for utility in real world applications and we focus on their effects at the levels of the population as a whole and of single individuals.

CCS CONCEPTS

• **Information systems → Spatial-temporal systems**; • **Security and privacy → Pseudonymity, anonymity and untraceability**; **Privacy-preserving protocols**.

KEYWORDS

Utility, geo-indistinguishability, *d*-privacy, anonymity, location based systems, mobility.

*The author contributed to this work while at LIX, École Polytechnique.

ACM Reference Format:
Adriano Di Luzio, Aline Carneiro Viana, Konstantinos Chatzikokolakis, Georgi Dikov, Catuscia Palamidessi, and Julinda Stefa. 2019. Catch Me If You Can: How Geo-indistinguishability Affects Utility in Mobility-based Geographic Datasets. In *Proceedings of* . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Many of the scientific challenges that we face today deal with improving the quality of our everyday lives. They aim at making the cities around us smarter, more efficient, and more sustainable. We study how to schedule public transport during peak hours, how to handle the traffic flow of commuters to and back from work, what is the most efficient path for waste disposal, the optimal locations for deploying charging stations for electric vehicles, and so on.

All these challenges share a common ground. They rely on datasets gathered from the real world that depict the mobility of hundreds of thousands individuals and picture, with great detail, the whereabouts of their lives—where they live, work, shop for groceries, and hangout with friends. These datasets are crucial to the scientific community as they provide researchers with the information they need to understand, capture, and model the human mobility and its patterns and, in turn, to solve the challenges that we face today.

At the same time, however, the collection of personal data also endangers the privacy of the users that to whom these data belong. To protect the privacy of the users, it is necessary to sanitize these datasets before releasing them to the public. In the case of mobility data, a well-known approach is *d*-privacy [4], which in the particular case of single locations is also known as geo-indistinguishability [2]. The main application of *d*-privacy is to protect from attacks such as stalking, inference of sensitive information associated with

exact locations, etc, but it can also be used to protect from re-identification attacks based on the adversary's knowledge of the mobility habits of an individual [15].

When we sanitize the datasets we trade the accuracy of the information they contain to protect the privacy of their users. The task of this paper is to shed light on the effects of this trade-off. We investigate on how sanitizing a mobile-based geographic datasets through d-privacy affects the quality of the information that it contains. In particular, we measure the quality of the sanitized Geolife dataset — *i.e.*, the utility of its information — by playing the role of a data scientist that leverages the dataset in two real world applications.

In the first application, we deploy an infrastructure of wireless hotspots in a wide geographic area (the city of Beijing) to offload the mobile cellular traffic of the users passing by. In the second application, we rely on the Geolife mobility traces to simulate a gossip-based opportunistic network.

These experiments capture the utility of the (sanitized) Geolife dataset from different points of views. When deploying the infrastructure of mobile hotspots, indeed, we investigate whether d-privacy maintains intact the key features of the dataset at the level of the entire population (*e.g.*, the statistical distribution of their movements) while protecting the specific whereabouts of each user. When we focus on the gossip-based opportunistic network, instead, we quantify the effects of d-privacy on the single individuals, measuring how distant — *i.e.* how unreliable — their mobility traces are when compared with those of the original dataset.

The rest of this paper is organized as follows. Next section discusses the rationale behind our work and considers the previous works related to differential privacy, geographical indistinguishability, and the utility metrics that we study. Section 3 illustrates the Geolife dataset, which will be used for our experiments. Section 4 explains how to apply d-privacy to obfuscate the data of Geolife. Section 5 show how to retrieve useful information from the obfuscated data for two applications: the placement of hotspots, and the gossip protocol. Section 6 discusses the effectiveness of our sanitization technique on Geolife. Section 7 concludes.

2 RATIONALE

A major source of concern about location privacy lies in the realization that with sufficiently accurate data, it is possible to precisely locate a user and track his movements throughout the day [6], giving rise to a variety of malicious activities such as robbing or stalking.

For instance, in Wisconsin there were episodes of men tracking women with GPS or other location devices [14]. In California, records from automatic toll booths on bridges were used in divorce proceedings to prove claims about suspicious movements of spouses [17]. The application "Girls Around Me", combined social media and location information to find nearby women who did not necessarily agree to be found, allowing to access their Facebook profiles with a single click [3]. Particularly worrisome is the perspective of potential combination with the users' most sensitive information, such as religious belief, political views, or sexual orientation.

The simple solution of anonymizing the location data, i.e. removing the name and any other personal identifier, is not effective. There are several studies that show, indeed, that *human mobility traces are highly unique*. For instance de Montjoye et al. [7], examined fifteen months of mobility traces generated by 1.5 million of individuals, users of a certain mobile phone operator. The experiments showed that 4 spatio-temporal points, randomly drawn from a trace, were enough to uniquely identify the trace in 95% of the cases. Song et al. [18] conducted similar experiments on a dataset of location-time data generated by about a million users over a period of a week, and showed that 2 points were enough to uniquely identify a trace in 60% of the cases. In combination with side knowledge (for instance, the approximate areas of domicile and work locations of an individual), the information contained in a trace can easily allow also the re-identification of the anonymized user.

Given the ineffectiveness of the anonymization technique, the scientific community has investigated other approaches. In particular, techniques based on obfuscation via controlled noise have emerged as a convincing alternative. Among these, *differential privacy* [10] and its distributed version, *local differential privacy* [9], have been particularly successful, and represent nowadays the cutting-edge of research on privacy protection.

Differential privacy (DP) was developed in the area of statistical databases, and it aims at protecting the individuals' data while allowing to make available the aggregate information. This is obtained by adding controlled noise to the query outcome, in such a way that the reported answer will not depend crucially on the data of a single individual. DP has also been used in the context of location privacy, see for example [5, 11, 12].

Local differential privacy (LDP) differs from DP in that users obfuscate their personal data by themselves, before sending them to the data collector. The advantage of LDP, with respect to DP, is that it does not need to assume a trusted third party, and since all stored records are individually-sanitized, there is no risk of privacy breaches due to malicious attacks.

In this paper we focus on d-privacy [4], which is a variant of LDP that can be applied whenever the space X of the data to be protected is provided with a notion of distance d. The idea behind d-privacy is to use the metric structure to achieve a better trade-off between privacy and utility. We say that

an obfuscation mechanism K is d-private if for every pair of secrets (data values to be protected) $x, x' \in X$, and for every measurable set S, we have $P[K(x) \in S] \leq e^{\epsilon d(x,x')} P[K(x') \in S]$, where $P[\cdot]$ represents the probability of an event and ϵ is a parameter representing the desired level of privacy. In other words, d-privacy allowing two data to become more and more distinguishable as their distance increases. Thus, it allows the adversary to infer some approximate information about the true value, but does not allow him to infer the exact true value. As explained in [4], d-privacy can be implemented by using an extended version of Laplace noise [1]. The instance of d-privacy to the case in which X is a set of geographical positions, and d is the geographical distance, is also known as *geo-indistinguishability* [2]. The version of the Laplace noise suitable for this case is called *planar Laplace*, and will be defined in Section 4. In this paper we focus on location data, and we will use d-privacy on the geographical data (geo-indistinguishability). In addition, we consider the removal of some intermediate points in a trace, aiming at reducing the degradation of privacy due to the combination of correlated information.

Concerning the utility, there are mainly two kinds: *quality of service* and *statistical information*. The first one refers to what the single user expects in exchange of his individual data. The second one refers to the accuracy of the information that we can extract from the collection of obfuscated data. In this paper we consider the second, and we analyze it in the context of two applications: the placement of hotspots, and the gossip protocol. In particular, in the first application the aim is to try to optimize the positioning of the hotspot, in such a way that we serve the highest number of users. In this scenario, it is clearly important to know the distribution of the user in the area of interest. Hence we should try to reconstruct the original distribution of the location data from the collection of the obfuscated ones, knowing the sanitization mechanism that has been used. In the second one, the goal is to study as accurately as possible how gossips propagate via the proximity of users, given that we have only obfuscated locations (and hence obfuscated proximity information) at our disposal.

In both cases, the utility depends crucially on the ground distance, which in our case is the geographical distance. The fact that the noise generated in d-privacy depends also on the distance allows to get a trade-off between utility and privacy superior to that of other LPD mechanisms.

A main contribution of our work is that we consider the trade-off between privacy and utility in some specific real-world application. Usually this trade-off is studied using an abstract notion of utility, like for instance the expected distance between real location and the corresponding obfuscated location [16].

[1]The standard Laplace distribution is defined on the real numbers, but as shown in [2] it can be extended to arbitrary metric spaces.

3 THE GEOLIFE DATASET

We evaluate our work against Microsoft's Geolife GPS trajectories dataset [19]. A trajectory in Geolife is a sequence of timestamped WGS84 latitude, longitude pairs with an associated user identifier. A set of volunteers, most of which working at Microsoft Research Asia, collected the trajectories by using a variety of GPS-loggers and GPS-enabled devices while commuting, going for shopping, and performing outdoor activities. In this work, we focus on the trajectories collected during the months of November and December 2008 and we only consider the geographic points that fall within the geographic center of Beijing, *i.e.* in the bounding box defined by the points of latitude $39.85 - 40.05$ and longitude $116.25 - 116.5$ (for an area of roughly $475\,\text{km}^2$). The resulting dataset consists of 39 users, 2075 trajectories, and more than 2 millions spatio-temporal points.

Figure 1 depicts a scatter plot of the latitude and longitude pairs in the dataset, highlighting the streets of Beijing's city-center and Microsoft Asia Headquarters in the top left of the figure. Each user contributed to the dataset to a different degree. On average, they recorded 50 000 latitude and longitude pairs (the standard deviation is 45 000). 6 users recorded less than 10 000 spatio-temporal points, while the 5 most active users contributed to the 35% of the dataset. The sampling rates and the collection times are also heterogeneous. Most of the trajectories have been collected with a sampling rate of 2 to 5 seconds. Nonetheless, some samples appear after occasional, longer gaps — almost a thousand pairs of consecutive geo-spatial points were collected more than 4 hours apart. Most of the volunteers commuted to the center of Beijing during the day (*e.g.* to reach Microsoft Research Asia headquarters) and then left at night. As a result, the majority of the samples was collected between 7 am and 9 pm.

4 PRIVACY

The Geolife dataset contains a sequence of spatio-temporal points labeled with the identifier of the user that collected them and with a trajectory number. Our goal is to protect the users' locations by sanitizing the dataset, *i.e.* by perturbing each spatio-temporal sample with a controlled amount of geographic noise that, at the same time, protects the exact location of the user while allowing to infer approximate information about it.

More precisely, we aim at protecting the original users' locations within a radius r with a level of privacy ϵ that depends on r. Towards this goal, we begin by focusing on how to perturb a single spatio-temporal sample $\boldsymbol{x}$ of the dataset. To do so, we define a planar Laplace distribution

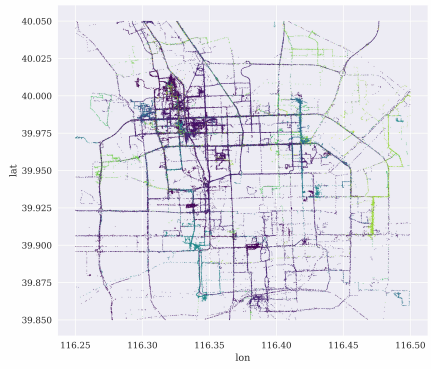

Figure 1: Geographic visualization of the trajectories in the Geolife dataset, depicting the streets of Beijing's city-center and Microsoft Asia Headquarters. The different colors corresponds to different users.

whose probability density function depends on ϵ (the desired privacy level for one unit of distance) and on the geographic coordinates of the point $\boldsymbol{x}$ (*i.e.* its WGS84 latitude and longitude pair). Then, we draw a random, new point $\boldsymbol{y}$ from the planar Laplace distribution, *i.e.* we add a controlled amount of Laplace noise to $\boldsymbol{x}$. The probability density function of the planar Laplace distribution of a point $\boldsymbol{x}$ according to ϵ on any other point $\boldsymbol{y}$ is defined as $D_\epsilon(\boldsymbol{x})(\boldsymbol{y}) = \frac{\epsilon^2}{2\pi} e^{-\epsilon \cdot ||\boldsymbol{x}-\boldsymbol{y}||}$, where $\frac{\epsilon^2}{2\pi}$ is a normalization factor and $||\boldsymbol{x}-\boldsymbol{y}||$ is the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$.

In practice, any point $\boldsymbol{y}$ can be represented, w.r.t. $\boldsymbol{x}$, as a pair (r, Θ), where $r = ||\boldsymbol{x} - \boldsymbol{y}||$ and Θ is the angle of the line connecting $\boldsymbol{x}$ to $\boldsymbol{y}$ with the respect to the horizontal axis of the reference coordinate system in use. In our case, we measure r as the Geodesic distance, in meters, between $\boldsymbol{x}$ and $\boldsymbol{y}$ and Θ as the azimuth of $\boldsymbol{y}$ in decimal degrees. For this reason, drawing a random point $\boldsymbol{y}$ from the planar Laplace distribution is equivalent to drawing, independently, r from a Gamma distribution of shape 2 and scale $1/\epsilon$ and Θ uniformly in $[0, 360)$. Then, perturbing $\boldsymbol{x}$ with the controlled Laplace noise is equivalent to solving a geodesic forward transformation, *i.e.* by obtaining a new pair of latitude, longitude coordinates by transforming $\boldsymbol{x}$ into a new point $\boldsymbol{y}$ at distance r from $\boldsymbol{x}$ and azimuth Θ.

Finally, we note that there is an inverse correlation between the level of privacy ϵ and the expected value of the Gamma distribution $\mathbb{E}[r] = \frac{2}{\epsilon}$. This value represents the average distance from the original location at which we perturb a spatial sample in the dataset, *i.e.*, the *expected noise* of the perturbation.

The next step consists of generalizing the perturbation to all the points of the dataset. In this work we define two different techniques. The first one deals with each spatio-temporal sample independently from all the others, by repeatedly sampling a new noisy point $\boldsymbol{y_i}$ from the planar Laplace distribution for each individual point $\boldsymbol{x_i}$ in the dataset. We call this technique *independent noise*. Figure 2 shows an example of this technique applied against a single Geolife trajectory: The top-left plot displays the un-sanitized trajectory (*i.e.* the real trace) and the top-right displays the same trajectory after perturbing, independently from each other, all the samples of the trajectory with 100 m of expected noise. The second technique, instead, also takes into account the temporal features of the dataset. More in details, let $w_{t_0:t_f}$ be single trajectory of a specific user, depicting her movements from time t_0 to time t_f. Our goal is to perturb the trajectory $w_{t_0:t_f}$ not only from the geographic dimension (as it happens with the independent noise) but also from its temporal dimension: *I.e.*, to provide an additional level of protection by hiding the movements of users across short periods of time. As such, we fix an interval of time T, *e.g.* of 5 minutes: Within that interval, we aim at protecting the real locations of the user and at hiding the frequency of her movements as well. To do so, we partition the original trajectory of the user into sub-trajectories that are non-overlapping, composed of adjacent points in the original dataset, and have a duration of at most T: Each sub-trajectory $w_{t_0:t_{T-1}}$, $w_{t_T:t_{2T-1}}$, etc., depicts her movements from times t_0 to t_{T-1}, from t_T to t_{2T-1}, and so on. We perturb the first spatio-temporal point $\boldsymbol{x_t}$ of each sub-trajectory (*i.e.* for t in t_0, t_T, etc.) by sampling a new noisy point $\boldsymbol{y_t}$ from the planar Laplace distribution and we replace all the remaining points of the sub-trajectory with $\boldsymbol{y_t}$. As a result, we obtain a set of sub-trajectories — each one composed of a perturbed point repeated over time. We call this technique *time-dependent noise* and the bottom-left plot of Figure 2 displays it in action with T set to 5 minutes and an expected noise of the perturbation of 100 m.

In both cases, we assume that the perturbation mechanism and its parameters (*i.e.*, the desired privacy level ϵ or, equivalently, the expected noise $\mathbb{E}[r]$) are public. For this reason, we also assume that an attacker would try to recover the real spatio-temporal points from the noisy trajectories by performing some additional processing on the dataset. As an example, with time-dependent noise we study an attacker that knows how each sub-trajectory is composed of the perturbed point at time $t_{k \cdot T}$ repeated over time. We assume that he is not naive, he is aware of this technique, and that he synthesizes the missing movements of the user within each sub-trajectory: *E.g.* by applying a geodesic-based linear interpolation to the samples between time $t_{k \cdot T}$ and $t_{(k+1) \cdot T}$, *i.e.* between the first noisy point of each sub-trajectory and the first of the next one. Analogously, we also consider an attacker that tries to cancel the Laplace noise applied to a

trajectory by leveraging its statistical properties. The trajectories in Geolife, indeed, have been collected by users on foot, on public transport, or on personal vehicles (*i.e.* cars or motorbikes). As such, the real points of a single trajectory lie on a smooth path from one geographic location into another. For this reason, when we perturb a large number of consecutive samples, the noisy spatio-temporal points could still reveal the shape of the original trajectory, as it happens in Figure 2. As a result, we also study an attacker that tries to remove the Laplace noise by considering, at the same time, a subset of noisy adjacent samples in a trajectory and aggregating their mean values by using a moving average: *I.e.*, selecting k adjacent samples, computing their mean, and then setting the resulting value as the $k+1$-th sample of the trajectory; next, discarding the oldest sample in the sequence, considering the $k+1$-th sample, and repeating the previous averaging steps to compute the $k+2$-th sample, and so on. We assume that an attacker would use such a technique both with independent noise and with time-dependent noise after applying geodesic linear interpolation to the sub-trajectories. The bottom-right plot of Figure 2 depicts the results obtained by applying geodesic linear interpolation to the noisy trajectory of the bottom-left plot, *i.e.* perturbed with the time-dependent noise technique at T set to 5 minutes and an expected noise of 100 m.

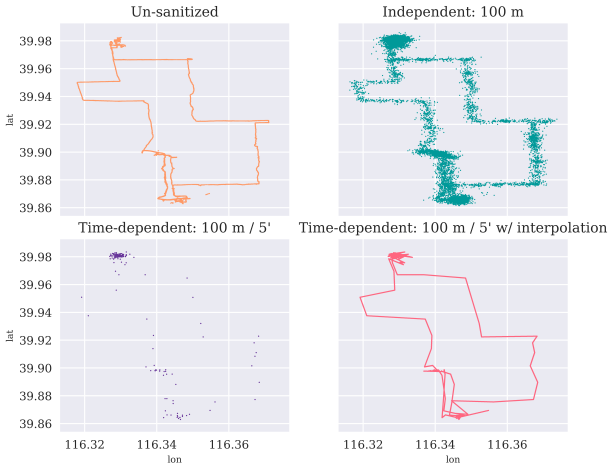

Figure 2: The effects of our sanitization techniques on a single Geolife trajectory (trajectory number 9053). The top-left plot displays the original trajectory; the top-right plot shows the effects of the independent noise technique with 100 m of expected noise; the bottom-left plot depicts the result of the time-dependent noise technique with 100 m of expected noise and $T = 5$ minutes; the bottom-right trajectory was obtained by applying geodesic linear interpolation to the trajectory of the bottom-left plot.

5 UTILITY

This paper sheds light on how sanitizing a mobility-based geographic dataset through d-privacy affects the utility of the underlying data. In particular, we focus on two different scenarios, inspired by two real world use cases that rely on the mobility of hundreds of thousands individuals. We base our first investigation on the work of Oliveira and Viana [13] that aims at deploying a network of mobile hotspots to offload cellular traffic through wireless technology. Here, we reproduce their experiment to compare the effectiveness of the hotspot networks deployed by relying on the original and sanitized versions of the Geolife dataset. While doing so, we aim at studying the effects of the d-privacy sanitization at the global scale: We study its effects on the population of Beijing as a whole and we focus on how it affects the statistical distribution of their movements. The second scenario, instead, tackles one of the classic problems in distributed systems and wireless networks: Gossip protocols [8]. The applications of gossip protocols allow modeling the spreading of epidemics, of data routing in ad-hoc networks, and of message passing in opportunistic networks. With this experiment, we study the effects of the d-privacy sanitization against the single individuals in our dataset. The reminder of this section describes in details the two investigations and their associated results.

5.1 Hotspots

Everyday, millions of devices (*e.g.* smartphones, tables, and Internet-of-Things enabled devices) leverage the mobile cellular network to communicate and exchange data. Cellular networks, however, only have a limited amount of bandwidth available and the quality of their communication decreases with the number of devices competing for it. To make things worse, the number of cellular-equipped devices and the amount of bandwidth that they require has increased exponentially over the last few years. As a result, we face the challenge of providing high-speed, low-cost, and ubiquitous mobile communication to these devices — in particular in densely crowded metropolitan areas. This experiment tackles this challenge.

We play the role of a mobile network operator that aims at decreasing the load on its cellular network by offloading its traffic through an infrastructure of WiFi hotspots deployed across the metropolitan area of Beijing. Their task is to study the mobility patterns of the users to decide where to deploy the WiFi hotspots: *I.e.*, to choose a minimum set of geographic positions for the hotspots that maximizes the expected amount of offloaded traffic. In practice, we fix the maximum range of radio communication between a user's smartphone and a WiFi hotspot to 50 m and, accordingly,

we tessellate the area of Beijing's city-center with a grid of squared cells of side $50\,\text{m} \cdot \sqrt{\pi}$. Each cell represents a possible position for an hotspot, has an area corresponding to the circle of radius 50 m, and simulates the wireless communication range. Overall, the tessellation results in 62230 cells. The goal of the mobile network operator is to select the subset of these cells that will host the hotspots by measuring the expected amount of traffic that a hotspot placed within each cell would offload. Typically, a network operator measure the expected traffic of a cell by taking into consideration, at the same time, the mobility of users, the amount of traffic that they generate by time of the day, and the geographic proximity of the cells to the key-locations of a metropolitan area (*e.g.*, the downtown district). For example, it considers the tendency of users to move within a confined environment and to repeatedly travel over the same paths, and favors cells that lie along these paths. In this work, however, we simplify the task of selecting which cells will host the infrastructure of hotspots. We assume that, at any time, each user is ready to offload a constant amount of traffic, that the offloading time is instantaneous, and that the offloading infrastructure has infinite bandwidth. As a consequence, every hotspot offloads an amount of traffic proportional to the number of users that move within its communication range at any time — more popular hotspots offload higher amounts of traffic. This allows us to select the cells based on the number of samples in the dataset that fall within their bounds, *i.e.* by their frequency distribution. In particular, we select the subset of cells according to their ranking in the frequency distribution and we favour more popular cells first.

The higher goal of the experiment is to study the extent to which perturbing the dataset with d-privacy affects the deployment of the hotspots. To evaluate its quality — the performance of the experiment — we measure the overall amount of traffic that the infrastructure offloads. Let k be the number of hotspots deployed, we define the score of the experiment p_k to be the percentage of traffic that they offload or, equivalently, the percentage of samples in the dataset that fall within the bounds of their cells. The unaltered Geolife dataset requires 19696 hotspots to offload all the traffic of its users and, for this reason, we set k to 19696 and we define $p_{19696} = 100\%$ to be the baseline score. We based the experiment on a perturbed dataset as follows. First, we consider the (noisy) distribution of users on the cells of the tessellation derived from their noisy mobility traces by counting the frequencies (*empirical noisy distribution*). Next, we select the $k = 19696$ cells that rank highest in the noisy frequency distribution, *i.e.* those cells that are most popular according to the samples of the perturbed dataset. Finally, we evaluate the amount of traffic that the infrastructure would offload when tested against the unaltered Geolife dataset.

Empirical noisy Distribution. In particular, we investigate the effects of the perturbations created by leveraging the independent noise methodology (Section 4) with values of expected noise between 50 m and 1000 m. Figure 3 depicts the scores of these experiments. The trend is inversely proportional: To higher values of expected noise correspond lower percentages of traffic offloaded by the infrastructure. Nonetheless, even when we perturb the dataset with an expected noise of 500 m or 1000 m, the resulting infrastructure offloads, respectively, more than 95% and almost 90% of the traffic. In fact, the amount of traffic offloaded decreases almost linearly with respect to the expected noise. In addition, these results demonstrate the effects of the perturbations on the distribution of the tessellated cells. When we perturb, on average, each sample 1000 m from its original position, the sanitized and unaltered dataset share 90% of the most popular cells. The left column of Figure 4 shows the phenomenon from a graphical point of view. There, each plot depicts the 2D frequency distribution of the cells resulting after the tessellation. To warmer colors correspond more popular cells and to transparent points correspond cells that were not visited by any user. From top to bottom, the plots show how perturbing the Geolife dataset with increasing values of expected noise affects the frequency distribution of the cells. In fact, they show how the distributions consistently highlight the most trafficked streets of Beijing and the headquarters of Microsoft Asia, regardless of the amount of expected noise applied to the mobility traces.

Iterative Bayesian update. As we have seen, we rank the cells of Beijing according to their frequency distribution; the results of the experiment depend exclusively on the geographic features of each dataset and on the mechanism, i.e., the kind and the amount of noise that we apply.

If the mechanism is public (*e.g.*, they are provided together with the sanitized dataset), then we can improve the result by trying to reconstruct as faithfully as possible the original distribution (i.e., the distribution obtained by counting the frequencies on the original data, before sanitization). We do so by leveraging the iterative Bayesian update (IBU) approach [1] that aims at recovering the original distribution from the noisy datasets, knowing the sanitization mechanism that has been used (in our case, the d-privacy noise). The IBU is a particular instance of the statistical *expectation-maximization method*, which results in a *maximum likelihood estimator*, namely a distribution that maximizes the probability of having produced, via the given sanitization mechanism, the observed noisy data.

Graphically, we aim at drawing a more focused heat-map of the cells in Beijing's metropolitan area and at improving, as a result, the performance of the hotspots deployed across the city when relying on a perturbed dataset.

Let $Y = y_1, y_2, \ldots, y_n$ be the perturbed dataset samples, *i.e.* the noisy pairs of latitude and longitude coordinates in the dataset. Let $\hat{q}$ be the estimated probability distribution of Y, that we compute by measuring the relative frequency of each coordinate pair. Let $Z = z_1, z_2, \ldots, z_m$ be the spatial samples before the perturbation (*i.e.*, every sample that, perturbed, could lead to any coordinate pair in Y) and let p be the probability distribution of Z. Let C be a stochastic matrix depicting the d-privacy method used to generate the noise, where $||z_i - y_j||$ depicts the geodesic distance between z_i and y_j: $C_{i,j} = P(Y = y_j|Z = z_i) = D_\epsilon(z_i)(y_j) = \frac{\epsilon^2}{2\pi} \cdot e^{-\epsilon \cdot ||z_i - y_j||}$. We estimate the probability distribution $\hat{p}$ of p (unknown) by leveraging the empirical distribution $\hat{q}$ derived from the perturbed dataset and the stochastic matrix C. The IBU proceeds towards its goal by iteratively refining a sequence of probability distributions $\hat{p}_1$, $\hat{p}_2$, $\hat{p}_3$, and so on, where $\hat{p}_1$ is a random distribution (*e.g.*, uniform) and $\hat{p}_{n+1}[z_i]$ with $z_i \in Z$ is defined as $\hat{p}_{n+1}[z_i] = \sum_{y_j \in Y} \hat{q}[y_j] \cdot \frac{C_{i,j} \cdot \hat{p}_n[z_i]}{\sum_{z_k \in Z} C_{k,j} \cdot \hat{p}_n[z_k]}$. The iterative process stops when $\hat{p}$ converges, *i.e.* when the absolute difference between $\hat{p}_{n+1}$ and $\hat{p}_n$ is smaller than a specified threshold δ: $\sum_{z_i \in Z} |\hat{p}_{n+1}[z_i] - \hat{p}_n[z_i]| < \delta$. When the process converges we set $\hat{p} = \hat{p}_{n+1}$. Then, we study the estimated probability distribution $\hat{p}$ of the possible positions within Beijing as the frequency distribution of cells and we select the $k = 1969$ most popular to deploy the hotspots.

The computational complexity of a single iteration of Bayesian update is linear in $|\hat{p}| \cdot |q|$ and depends, respectively, on the area of the geographic region under examination and on the number of samples in the dataset. In our case, $\hat{p}$ encodes the possible positions within the city center of Beijing and contains millions of elements; in turn, the IBU has a significant complexity. We implemented the iterative process through the Nvidia CUDA Programming API and then performed the computation on a cluster of GPUs. The number of rounds required by $\hat{p}$ to converge depends on the expected noise applied to the dataset and on the value of δ. When fixing δ to 10^{-8} the process converged, on average, after 50 iterations.

The green line of Figure 3 depicts the effects of the IBU on the perturbed dataset. As before, the scores decrease almost linearly with respect to the expected noise applied to the dataset. Nonetheless, the IBU improves the performance of the hotspots deployed, on average, by 2 percentage points — *e.g.*, improving the score at 250 m of expected noise from 96% to 98%. The effects of the IBU are even more striking when we consider stronger perturbations of the dataset. At 1000 m of expected noise the score increases to 93% from 89%. The right column of Figure 4 depicts a graphical representation of the results. There, each heat-map shows how the IBU refines the frequency distribution on the cells, by starting from the perturbed datasets of the corresponding heat-map on the left.

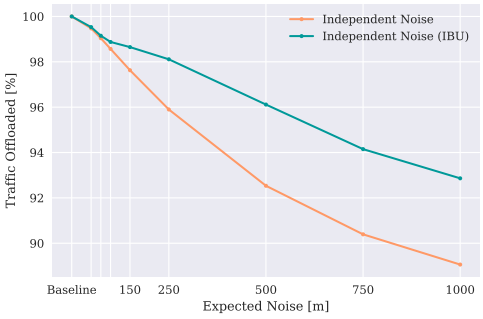

Figure 3: The percentage of traffic offloaded against the expected noise used to sanitize the Geolife dataset with the independent noise mechanism. The green line depicts the results of the sanitized Geolife datasets after the application of the Bayesian Update method.

The process effectively decreases the noise in the heat-maps, focusing the points with more hits closer to the neuralgic areas of the city — for example, around its most trafficked streets and Microsoft's headquarters. Graphically, the perturbation applies a convolution on the samples of the dataset that blurs the heat-maps of the frequency distributions across the cells of Beijing. The IBU (partially) recovers the original focus of the heat-maps and of the corresponding frequency distributions and, in turn, improves the deployment of the hotspots.

These results have multiple implications. First, they show how the sanitization of mobility-based datasets through geo-indistinguishability affects the underlying data at a global scale: Their statistical distributions remain comparably similar to the originals, even at high levels of expected noise. In turn, in use cases based on statistical features, they show how the sanitization process preserves the privacy of the users and, at the same time, does not jeopardize the utility of the dataset. Lastly, they show how leveraging a statistical update technique, such as the IBU, can further bridge the gap between the original and the perturbed statistical features — increasing the utility of the datasets without reducing the privacy of the users.

5.2 Gossip

To continue our investigation on the trade-offs between privacy and utility in mobility-based datasets, we turn our attention to a gossip-based communication protocol [8]. This protocol draws its inspiration from how epidemics spread in nature, and is simple, fast, and scales efficiently in the size of the population. In addition, the execution of the protocol models an opportunistic network where two users can only

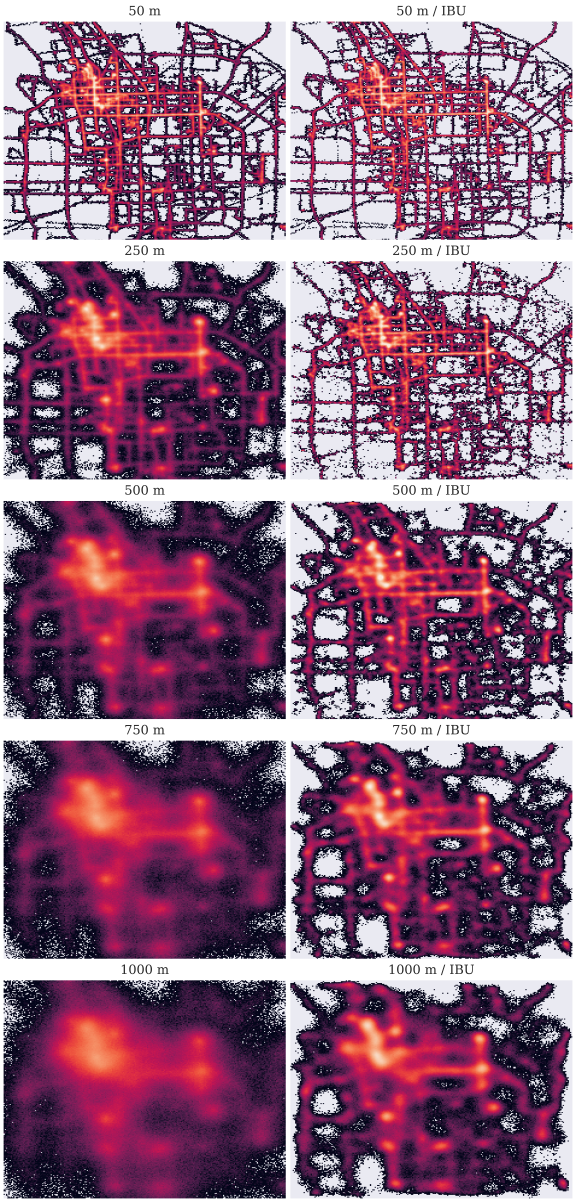

Figure 4: The heat-maps depict the frequency distribution of the cells resulting from the tessellation of Beijing's metropolitan area, *i.e.*, the number of users that crossed the boundary of each cell. Warmer colors correspond to the more popular cells and empty points correspond to the cells that were not visited by any user. The left column depicts the heat-maps resulting from the perturbation of the datasets at increasing values of expected noise. The right column depicts the heat-maps resulting from post-processing the noisy datasets of the left column with the iterative Bayesian update process.

exchange messages based on their physical proximity to one another. Here, we rely on the Geolife dataset to simulate the opportunistic network, and we play the role of a network engineer that studies it to measure the infection rate (*i.e.*, the percentage of users reached by the message) and its latency.

The gossip protocol begins by selecting a random source among the Geolife users to broadcast a single message to all the others. We discretize time into slots of 5 seconds. Two users can communicate if any of their spatio-temporal samples appear within 50 m from one another at the same time slot. The protocol ends when either the message reaches all users or the simulation exhausts all mobility traces.

The goal of this experiment is to understand to which extent perturbing the dataset with d-privacy affects the simulation. In particular, we aim at discovering if there exists a perturbation that protects the privacy of the users while also resulting in a realistic simulation, *i.e.* one that mimics the execution of the protocol against the unaltered Geolife dataset. To reach our goal we apply different perturbations to the dataset and we measure the resulting infection rate. We repeat every simulation 1000 times, each time selecting as new random gossip source, and we record the average infection rate. When we base the execution of the protocol on the unaltered Geolife dataset, the message reaches 40% of the users on average — this is the baseline of the experiment. In this case the message reaches less than half the population because not all users contributed to the Geolife dataset to the same extent (cf. Section 3) and some individuals participated to the data collection for less than a day worth of samples: These individuals are less likely to be reached by the gossip. We compare the baseline results with the simulation of the protocol after perturbing the dataset at increasing levels of expected noise and we study the effects of both the independent and time-dependent mechanisms (with geodesic linear interpolation, as described in Section 4).

Figure 5 depicts the results of the experiment. The horizontal axis describes the amount of expected noise that we use to perturb the dataset; the baseline corresponds to the unaltered Geolife dataset. The vertical axis measures the average infection rate across the 1000 repetitions of each simulation. Finally, the different lines depict the different perturbation techniques, *i.e.*, the independent noise and time-dependent noise at increasing values of T. We begin by noting that each line exhibits an increasing trend: To stronger sanitization (*i.e.*, higher values of expected noise) corresponds higher average infection rate. The Laplace noise that perturbs the traces, indeed, scatters the positions of the users around the city (Figure 2 provides a graphical representation of the phenomenon). Two users that did not meet in the unaltered Geolife dataset are now more likely to infect each other and, in turn, to trigger a chain reaction that will eventually infect most of the population. The stronger the

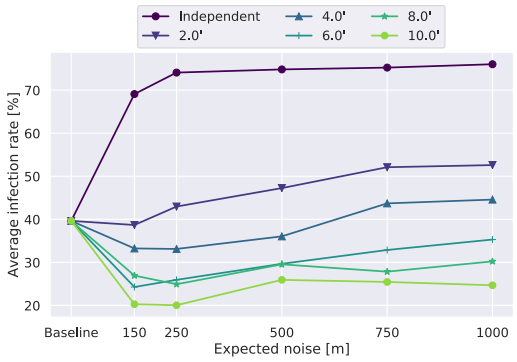

Figure 5: The percentage of users reached, on average, by the 1000 simulations of the gossip protocol while sanitizing the dataset at increasing levels of expected noise. The different lines correspond to the different sanitization techniques. The baseline depicts the unaltered Geolife dataset.

Laplace noise, the higher the scattering of users across the city, and the more chances for the infection to happen. The independent noise mechanism, in particular, results in the highest overall infection rate: With this mechanism, indeed, each user counts as many samples as in the unaltered dataset, but her whereabouts result scattered in a larger fraction of the metropolitan area. The sanitization based on time-dependent noise, instead, shows lower infection rates. This techniques effectively reduces the freedom of movement of each user by sampling their noisy position only once for each period of time considered, and then keeping them stationary until the next time frame. For this reason, to higher values of T considered correspond lower infection rates. In particular, when we allow users to move once every ten minutes the infection rate drops to half the baseline value, even when considering an adversary that performs linear interpolation of the intermediate positions. Finally, we note how smaller values of T result in a good balance between privacy and utility. When sampling time every 2 minutes, for example, the utility of the dataset remains comparable to the baseline for values of expected noise lower than 250 m. Next, it increases almost linearly with respect to the expected noise sampled. Even when applying our strongest sanitization (1000 m) the average infection rate results in roughly half of the population: 10% more than the baseline result. As a result, our sanitization methodology based on small time windows enables us to protect the privacy of the users in the dataset while, at the same time, studying a gossip protocol that produces a simulation comparable to the baseline.

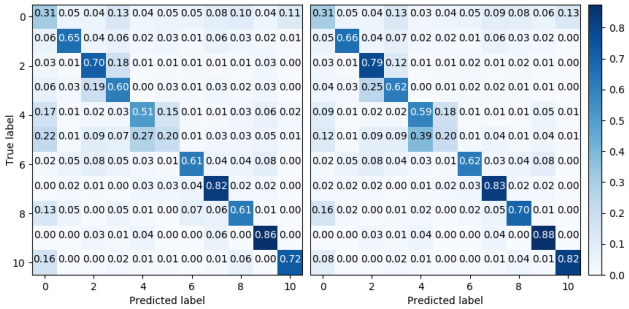

Figure 6: Confusion matrices of the de-anonymizer; time-independent (left) and time-dependent (right) noise.

6 ON d-PRIVACY FOR ANONYMITY

We finish with a discussion emphasizing the challenges of location obfuscation when the privacy goal at hand is that of *anonymity*. Perturbation techniques such as d-privacy were designed in the context of Location Based Services, to prevent the service provider from learning the user's location with accuracy. Indeed, geo-indistinguishability requires that locations close to each other produce the same output, thus they cannot be distinguished by the adversary. This is made explicit by a Bayesian characterization of this notion [2], roughly stating that the adversary's knowledge of the user's position, within a small radius from the true location, is improved only negligibly by observing the noisy location. Note that the service provider is typically assumed to know the user's identity: For instance the user might be logged-in to get personalized recommendations.

However, sometimes the user's goal is anonymity, *i.e.*, not hiding his location, but his identity. In this context, the provider has prior information on the mobility patterns of a group of users, and Alice's behaviour should not be distinguishable within this group. Although perturbing her location could be beneficial, Alice's anonymity highly depends on the behaviour of the other users in the group. The fact that the adversary cannot know Alice's location accurately might be useless, if Alice is the only user within a large radius.

To demonstrate this fact we performed a de-anonymization attack on a sanitized version of the Geolife dataset. We randomly selected 10 target users and used 11 classes for the classifier: one for each target user, and one (class 0) for all the others together. We used a Recurrent Neural Network with 3 LSTM layers and 1 fully connected layer with ReLU activation, all regularised by a dropout rate of 0.1. The network was trained on obfuscated traces with an expected noise of 100 m ($\epsilon = 0.02$), using both time-independent and time-dependent (T=5 mins) variants.

The confusion matrices of the classifier are shown in Figure 6; we can see that the network can effectively de-anonymize obfuscated traces with an accuracy of 47.27% in the time-independent and 48.76% for the time-dependent case. This demonstrates that within the Geolife dataset, the number of users is too small and their behaviour too unique to effectively prevent the adversary from de-anonymizing them, even in the presence of noise.

7 CONCLUSIONS

In this work we investigated on whether it is possible to sanitize a mobility-based geographic dataset in such a way that it safeguards the privacy of users and that, at the same time, it preserves the quality of the underlying information. To reach our goal, we first developed a set of sanitization techniques based on d-privacy that protect the users by perturbing their mobility traces. Then, we measured how these techniques affect the utility of the dataset through two different experiments, based on the deployment of hotspots and on gossip protocols, that respectively targeted the statistical distributions of the geographical samples at a global scale and the whereabouts of single individuals at the local level. We found out that d-privacy preserves the statistical distribution of the dataset even at high privacy regimes and, as a result, is highly effective when studying the users' population as a whole. On the other hand, when we focused on the local level and on the whereabouts of the single individuals, the trade-off between privacy and utility appeared more marked. In our case, stronger privacy-preserving sanitization corresponded to less reliable simulations of a gossip protocol and a sanitization that preserves both privacy and utility required more careful considerations. Lastly, we hope that this work will encourage more individuals and organizations to sanitize and then share their geographic datasets. Even after protecting the privacy of users, the sanitized datasets preserve the statistical properties of the original data and, in some cases, also the quality of the information at the local level. As a result, these datasets are crucial to the research community and could bring us one step closer to solving the challenges that we face every day.

REFERENCES

[1] Agrawal, R., Srikant, R., and Thomas, D. Privacy preserving OLAP. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2005), SIGMOD '05, ACM, pp. 251–262.

[2] Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., and Palamidessi, C. Geo-indistinguishability: differential privacy for location-based systems. In *Proc. of CCS* (2013), ACM, pp. 901–914.

[3] Brownlee, J. This Creepy App Isn't Just Stalking Women Without Their Knowledge, It's A Wake-Up Call About Facebook Privacy (Update), 2012. http://www.cultofmac.com/157641/.

[4] Chatzikokolakis, K., Andrés, M. E., Bordenabe, N. E., and Palamidessi, C. Broadening the scope of Differential Privacy using metrics. In *Proc. of PETS* (2013), vol. 7981 of *LNCS*, Springer, pp. 82–102.

[5] Chen, R., Ács, G., and Castelluccia, C. Differentially private sequential data publication via variable-length n-grams. In *Proc. of CCS* (2012), ACM, pp. 638–649.

[6] Cheung, A. Location privacy: The challenges of mobile service devices. *Computer Law & Security Review 30*, 1 (2014), 41–54.

[7] de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. Unique in the crowd: The privacy bounds of human mobility. *Nature Scientific Reports 3*, 1376 (2013).

[8] Demers, A., Greene, D., Hauser, C., Irish, W., Larson, J., Shenker, S., Sturgis, H., Swinehart, D., and Terry, D. Epidemic algorithms for replicated database maintenance. In *Proceedings of the Sixth Annual ACM Symposium on Principles of Distributed Computing* (New York, NY, USA, 1987), PODC '87, ACM, pp. 1–12.

[9] Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *Proc. of FOCS* (2013), IEEE Computer Society, pp. 429–438.

[10] Dwork, C., Mcsherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proc. of TCC* (2006), vol. 3876 of *LNCS*, Springer, pp. 265–284.

[11] Ho, S.-S., and Ruan, S. Differential privacy for location pattern mining. In *Proc. of SPRINGL* (2011), ACM, pp. 17–24.

[12] Machanavajjhala, A., Kifer, D., Abowd, J. M., Gehrke, J., and Vilhuber, L. Privacy: Theory meets practice on the map. In *Proc. of ICDE* (2008), IEEE, pp. 277–286.

[13] Oliveira, E. M. R., and Viana, A. C. From routine to network deployment for data offloading in metropolitan areas. In *2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)* (June 2014), pp. 126–134.

[14] Orland, K. Stalker Victims Should Check For GPS. The Associated Press, 2003. http://www.cbsnews.com/news/stalker-victims-should-check-for-gps/.

[15] Romanelli, M., Palamidessi, C., and Chatzikokolakis, K. Generating optimal privacy-protection mechanisms via machine learning. *CoRR abs/1904.01059* (2019).

[16] Shokri, R., Theodorakopoulos, G., Boudec, J.-Y. L., and Hubaux, J.-P. Quantifying location privacy. In *Proc. of S&P* (2011), IEEE, pp. 247–262.

[17] Simerman, J. FasTrak to courthouse. East Bay Times, 2007. http://www.eastbaytimes.com/2007/06/05/fastrak-to-courthouse/.

[18] Song, Y., Dahlmeier, D., and Bressan, S. Not so unique in the crowd: a simple and effective algorithm for anonymizing location data. In *Proceeding of the 1st Int. Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security* (2014), vol. 1225 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 19–24.

[19] Zheng, Y., Li, Q., Chen, Y., Xie, X., and Ma, W.-Y. Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing - UbiComp 08* (2008), ACM Press.

[20] Zheng, Y., Xie, X., and Ma, W. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull. 33*, 2 (2010), 32–39.

[21] Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web* (New York, NY, USA, 2009), WWW '09, ACM, pp. 791–800.